动 物 学 研 究 2005, Oct. 26 (5): 555-564 CN 53-1040/Q ISSN 0254-5853
Zoological Research

串联重复序列的物种差异及其生物功能

高 焕[1,2,3], 孔 杰[1,*]
(1. 中国水产科学研究院黄海水产研究所 农业部海洋渔业资源可持续利用重点开放实验室, 山东 青岛 266071;
2. 中国科学院海洋研究所, 山东 青岛 266071; 3. 中国科学院研究生院, 北京 100039)

摘要: 串联重复序列是指1~200个碱基左右的核心重复单位, 以头尾相串联的方式重复多次所组成的重复序列。它广泛存在于真核生物和一些原核生物的基因组中, 并表现出种属、碱基组成等的特异性。在基因组整体水平上, 各种优势的重复序列类型不同。即使在同一重复序列类型内部, 不同重复拷贝类别 (如AT、AC等) 在基因组中的存在也表现出很大的差异。同时, 这些重复序列类型和各重复拷贝类别在同一物种的不同染色体间, 以及基因的编码区和非编码区间也表现种属和碱基组成差异。这些差异显示了重复序列起源和进化的复杂性, 可能涉及到多种机制和因素, 并与生物功能密切相关。另外, 由于重复序列分析软件和统计标准还存在算法、重复长度、完美性等问题, 需要进一步探讨。此外, 串联重复序列的自身进化关系、全基因组水平上的进化地位、在基因组中的生物功能、重复序列数据库建立和应用研究等, 将是今后研究的主要课题。

关键词: 串联重复序列; 微卫星; 小卫星; 基因组; 起源与进化
中图分类号: Q31; Q75; Q819 **文献标识码**: A **文章编号**: 0254-5853 (2005) 05-0555-010

Distribution Characteristics and Biological Function of Tandem Repeat Sequences in the Genomes of Different Organisms

GAO Huan[1,2,3], KONG Jie[1,*]
(1. Key Laboratory for Sustainable Utilization of Marine Fisheries Resources, Ministry of Agriculture, Yellow Sea Fisheries Research Institute, the Chinese Academy of Fishery Sciences, Qingdao 266071, China; 2. Institute of Oceanology, the Chinese Academy of Sciences, Qingdao 266071, China; 3. Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)

Abstract: Tandem repeat sequences, also known as direct repeats, are repeat sequences in which the length of the repeat unit changes mainly from 1 to 200 bp size, and the repeat unit is arranged in a "head-tail" conjunction mode, and is distributed widely in the genome of eukaryotes and some prokaryotes. At the level of full genomes, both the abundance and distribution characteristics of repeat types, such as dinucleotide repeats and trinucleotide repeats et cetera are varied in different organisms, and the variedness also occurs in different repeat classes, such as AT and AC repeat classes etc. and across inter-chromosomes, and even between coding regions and noncoding regions. All of the above differences indicate that the genesis and evolution of tandem repeat sequences are complex and may involve several mechanisms and factors, as is typical of biology. Additionally, there exist some problems preventing us from further studying the tandem repeat sequences, e.g. the software to analyze repeat sequences, criteria such as the length, the copy number, and the perfect or imperfect delimitation to determine what is a repeat sequence or not which varies across researchers. In order to address these problems, six future research directions should be pursued: The study of tandem repeat sequences, the self-evolution relations of tandem repeat sequences, the evolution status in the level of full genomes, the biology function, the establishment of tandem repeat sequence data-banks, and their application researches.

Key words: Tandem repeat sequences; Microsatellites; Minisatellites; Genome; Genesis and evolution

---

收稿日期 2005-04-07; 接受日期: 2005-07-14
基金项目: 国家重点基础研究发展规划 (973) (G1999012007); 国家高技术研究发展计划 (863计划) (2003AA603021); 中国水产科学院水产种质资源与养殖技术重点开放实验室开放基金 (南海水产研究所)
* 通讯作者 (Corresponding author), E-mail: kongjie@sina.com

微卫星（microsatellites）重复序列是一类目前应用广泛的遗传标记，其重复单位碱基数目一般为1~6个，如（CA）n、（CAG）n等（Gao et al，2004），而与之相应的是小卫星（minisatellites）重复序列，其重复单位的碱基数目在不同文献（Vergand & Denoed，2000；Ingavale et al，1998；Jauert et al，2002）中有所不同。综合这些文献，小卫星重复单位的长度应定义在7~200个碱基，而有时又把其中25个碱基以上的重复单位所组成的重复序列称为大卫星（macrosatellites）（Wickstead et al，2003）。微卫星和小卫星重复序列因其核心重复单位是以头尾相连的多次重复的碱基组成，有别于回文序列（palindromic sequences）和反向重复序列（reversed repeat sequences），故统称它们为串联重复序列（tandem repeat sequences）。

在微卫星和小卫星重复序列的研究中，最为人们所熟知的是其作为分子遗传标记的研究。微卫星重复序列在群体间和不同个体间通常表现出很高的序列变异性，并且这种变异呈现共显性遗传，因而微卫星重复序列广泛应用于遗传多样性分析（Hadonou et al，2004；Romero et al，2003）、连锁图谱制作（Staten et al，2004）、疾病连锁分析（Sakurai et al，2004）和家系标识（Selvamani et al，2001）等研究。而小卫星重复序列常被制作成DNA探针，以基因组杂交的方式研究其DNA指纹图谱的特征（Jeffreys et al，1985；Saha & Bamezai，2000）。目前关于微卫星和小卫星重复序列的相关研究进展很快，每年都有数千篇研究成果文献的报道。一些综述性文献也对此进行了阐述（Luo et al，2003；He，1998），但对于这些串联重复序列在各物种基因组中的存在状况，以及这些存在状况与重复序列的起源和进化关系等方面都还缺少系统详尽地阐释。近年来，随着各种生物基因组测序计划的进行，许多科学工作者开展了基于基因组整体水平的串联重复序列的分析工作。

1 串联重复序列的差异

1.1 物种差异

目前用于生物基因组中重复序列分析的序列主要来源于已经公布的核酸数据库和各研究单位构建的随机DNA基因组文库，前者即美国的国家生物技术信息中心（National Center for Biotechnology Information，NCBI）、欧洲分子生物学实验室、日本国立遗传学研究所共同制作的国际核酸序列数据库（DDBJ/EMBL/GENBANK），其优点是全面系统，但只能局限于人类、拟南芥等少数已经完成基因组测序计划的物种；后者虽然不全面，但类似于从大群体随机抽样的方法，通过随机基因组克隆序列分析，可以得知该物种基因组中的串联重复序列的存在状况。

串联重复序列在部分原核生物的基因组中业已存在；在高等生物的基因组中更是比比皆是，如6个碱基及其以上长度的串联重复序列约占大肠杆菌（*E.coli*）基因组序列总长度（109 kb）的2.4%（Gur-Arie et al，2000）；在人类基因组中，重复单位在1~11 bp范围内的串联重复序列约占基因组长度的2%（Borstnik & Pumpernik，2002）。一般，按组成重复单位的碱基数目，串联重复序列分为单碱基、两碱基、三碱基等类型。而每一种重复序列类型又可细分出不同的重复拷贝类别，如CT和CG分别属于不同的重复拷贝类别（Gao et al，2004）。不同生物基因组中占优势的重复序列类型既相同，又不同（表1）。在原核生物和酵母的基因组中，处于优势的重复序列类型是三碱基，而比它们更高等的生物基因组中，则倾向于两碱基和单碱基重复序列类型。

重复拷贝类别也因种而异。部分生物基因组中占优势的重复拷贝类别的情况见表2。在现已研究的所有物种的单碱基重复序列中，A或T重复拷贝最多，而C或G很少；两碱基重复序列中，较低等生物基因组中重复拷贝以AT和AG为主，而高等生物基因组中又以AC最多。三碱基及其以上重复序列中，各种重复拷贝类别无明显规律可循，但最明显的特征是，处于优势数目的重复拷贝类别都富含A或T。

1.2 在染色体上的差异

重复序列各种类型及类别不仅在物种间不同，而且在同一物种的不同染色体上也不相同。如：果蝇微卫星重复序列在X染色体上的密度比在其他染色体上高得多（Bachtrog et al，1999）；按蚊微卫星重复序列在X染色体上的密度也是最高的，其平均长度是142.75 bp/Mb（Yu et al，2005）。人类的21和22号染色体的长度几乎相同，但重复序列的单碱基、三碱基和四碱基类型在22号染色体上的丰

表 1　不同物种基因组的优势重复序列类型

Tab. 1　Predominant repeat types in the genomes of different organisms

重复序列类型 Repeat sequences type	物种 Species	资料来源 Source of data
三碱基 Trinucleotide	生殖支原体 *M. genitalium*	Klevytska et al, 2001
	耶尔森氏菌 *Y. pestis*	Belkum et al, 1998
	酵母 *Saccharomyces cerevisiae*	Katti et al, 2001
两碱基 Dinucleotide	果蝇 *Drosophila* sp.	Ross et al, 2003
	哺乳类 Mammalia 脊椎动物 Vertebrate 节肢动物 Arthropod 啮齿类 Rodent	Tóth et al, 2000
单碱基 Mononucleotide	秀丽隐杆线虫 *Caenorhabditis elegans* 拟南芥 *Arabidopsis thaliana* 人类 Human	Katti et al, 2001
	有胚植物 Embryophyta	Tóth et al, 2000

表 2　不同物种基因组的重复拷贝类别

Tab. 2　Predominant repeat classes in the genomes of different organisms

物种 Species	两碱基 Dinucleotide			三碱基 Trinucleotide			四碱基 Tetranucleotide			五碱基 Pentanucleotide	六碱基 Hexanucleotide	资料来源 Source of data
	1	2	3	1	2	3	1	2	3	1	1	
拟南芥 *A. thaliana*	AT	AG	AC	AAG	ATG	AAC						Katti et al, 2001
大肠杆菌 *E. coli*	CG	AG	AC				GCCA					Gur-Arie et al, 2000
酵母 *Saccharomyces cerevisiae*	AT	AC	AG	AAC	AAT	AAG						Katti et al, 2001
秀丽隐杆线虫 *C. elegans*	AG	AC	AT	AAG	AAT	ATG						
中国明对虾 *F. chinensis*	AT	AC	AG	AAT	AAG	ATC	AGAT	ACAT	AGAC	AGAGA	ATTATC	Gao et al, 2004
家蚕 Silkworm	AG	AT	AC	AAT	AGC	AAG	AATC	AATG	AACC	AAAAN	AAAAAN	Li B et al, 2004
蚊子 Mosquito	AC	AG	AT	AGC	AAC	ACC	AATC	AATG	AACC	AAAAN	AAAAAN	
果蝇 *D. arizonae*	AC	AG	AT	AGC	AAC	ATC	ACAG			AGCTC	AACAGC	Ross et al, 2003
斑马鱼 Zebrafish	AC	AG	AT	AAG	AGC	AGG	AATC	AATG	AACC	AAAAN	AAAAAN	Li B et al, 2004
河豚 *F. rubripes*	AC	AG	AT	AGG	AGC	AAT	AAAT	ACAG	ACGC	AAAAN	TTAGGG	Edwards et al, 1998
家鼠 Mouse	AC	AG	AT	AGC	AAC	ACC	AATC	AATG	AACC	AAAAN	AAAAAN	Li B et al, 2004
小鼠 Rat	AC	AG					AAAN	AAGG				Beckman & Weber, 1992
人类 Huaman	AC	AG		AAN			AAAN					

富度明显大于 21 号染色体；而两碱基正好与此相反（Katti et al, 2001）。此外，重复拷贝类别在不同染色体上的密度也有较大差异：人类 GATA 重复拷贝类别在 Y 染色体中的密度达 222 bp/Mb，比在其他染色体上高得多（Subramanian et al, 2003）。

1.3　在基因和基因间隔区上的差异

重复序列在基因编码区（外显子）和非编码区（内含子和基因间隔区）同样存在差异。在原核生物弗氏志贺菌（*Shigella flexneri*）中，三碱基、四碱基和六碱基在编码区较多，而单碱基和两碱基则在非编码区较多（Yang et al, 2003）。与此不同的是，两碱基、三碱基、四碱基、五碱基、六碱基在大肠杆菌（*E. coli*）基因组的编码区和非编码区无显著差异，但单碱基却因统计重复序列长度的起点不同，而显示出有差异，如在开放阅读框（open reading frames, ORFs）区域（约占全基因组的 79.5%）和非编码区（约占全基因组的 20.5%），3 个碱基以上长度（即只要连续出现 3 个相同碱基或以上，就把此序列作为单碱基重复序列来统计）的单碱基重复序列的比例分别为 78.0% 和 22.0%，与编码区和非编码区的分布规律基本一致；而当把统计单碱基重复序列长度的标准逐渐提高的情况

下，如分别在4、5、6、7、8等碱基长度的情况下，单碱基重复序列的数量在开放阅读框（编码区）中的比例明显下降，而主要集中在非编码区的序列中（Gur-Arie et al，2000）。

在真核生物的基因组中，重复序列很少与编码区相连，而主要位于基因以外的区域（Cox & Mirkin，1997）。在灵长类、哺乳类、啮齿类、节肢动物、秀丽隐杆线虫、有胚植物、酵母和其他真菌的基因组中，在外显子区域最多是三碱基，其次是六碱基。同时，除了有胚植物和秀丽隐杆线虫外，其余在内含子和基因间隔区，六碱基的数量都比在外显子中多。灵长类内含子和基因间隔区中单碱基的数量最多，是两碱基和四碱基的两倍以上。在啮齿类、哺乳类、节肢动物和秀丽隐杆线虫非编码区的内含子和基因间隔区中最多是两碱基。值得一提的是，脊椎动物基因组非编码区中的四碱基比三碱基多；而在非脊椎动物和真菌中，四碱基则很少（Tóth et al，2000）。

重复拷贝类别在编码区和非编码区也有显著差异。在单碱基重复序列类型中，A主要存在于非编码区，C主要存在于编码区。两碱基中，AG在秀丽隐杆线虫的基因间隔区最多，AT则在内含子中最多；AC在真菌基因组的内含子中较多；AT是其他真核生物基因组的非编码区中最为丰富的类别。三碱基中，G+C丰富的重复拷贝类别在所有脊椎动物基因组的外显子中都是最多的；AAC和AAG在有胚植物外显子中最丰富；而A+T丰富的三碱基重复拷贝在酵母和真菌外显子中最为丰富。大多数物种基因组的外显子中一般都较少含有四碱基重复序列类型，但AAAB（B为除A之外的碱基）在灵长类和啮齿类中是最丰富的。五碱基重复序列类型主要存在于非编码区，其中主要是A+T丰富类型（Tóth et al，2000）。还有一类被称为编码氨基酸的重复序列，它们是编码蛋白质中的连续氨基酸（如丙氨酸-丙氨酸-丙氨酸）的重复序列，在这些区域的三碱基重复序列中，明显是倾向于G+C丰富类型的。在人类中最多的类别是CAG和GAG（Alba & Guigo，2004）。

2 串联重复序列的变异

串联重复序列在生物机体内很不稳定，经常发生变异，即重复单位的扩展或缩小。微卫星突变率因物种不同而不同。从果蝇中的 5×10^{-6}（Vazquez et al，2003），到人类中的 1×10^{-3}（Brinkmann et al，1998；Xu et al，2000）不等，这也说明重复序列是生物基因组DNA进化的一个重要来源。对于串联重复序列的起源及其进化机制一直都是研究的热点。较早对这种变异做出的解释是：当DNA复制的时候，会产生DNA滑移错配，随后，这种错配在DNA的复制过程中，经过修复和重组作用而产生重复的序列（Levinson & Gutman，1987）。但从各种重复序列类型在不同生物基因组间、同一基因组内不同染色体间、编码区和非编码区间存在的差异来看，不是某一个单一的机制就可以完全解释清楚这些差异的。

2.1 在基因组中的扩展机制

重复序列与许多疾病的发生存在着密切的联系。在约30%的结肠肿瘤患者中，肿瘤细胞与正常的健康细胞相比，其DNA序列中（CA）n重复序列在长度上有着显著的差异（Thibodeau et al，1993）。三碱基串联重复序列至少与16种遗传疾病的发生有关（Margolis & Ross，2001）。其中，(CAG)n/（CTG)n与亨廷顿舞蹈症（Huntington's disease）和强直性肌萎缩症（myotonic dystrophy）；(GCC)n /（GGC)n与脆性X染色体综合征（fragile X syndrome）；（GAA)n/（TTC)n与弗里德赖希共济失调（Friedreich's ataxia，FRDA）的发病有关。其发病的机理是由于位于基因编码区的这些三碱基重复序列的扩展造成的（Heidenfelder et al，2003）。Heidenfelder et al（2003）利用聚合酶在体外对GAA/TTC的重复序列进行扩增，并用电子显微镜观察其二级结构，首次发现了在DNA合成过程中(GAA)n和（TTC)n所形成的发夹环结构。这种发夹环结构的形成，可以保护后滞链滑动所形成的重复结构免受错配修复系统的修正。这也是支持重复序列发生的“复制滑移学说”的一个证据。

Heidenfelder & Topal（2003）进一步的研究表明，这种发夹环结构的形成并非是DNA复制过程中重复序列扩展所必需的。在Ⅱ型强直性肌萎缩症(myotonic dystrophy type 2，DM2）基因中的某些重复序列扩展位点上，并行存在着几种类似的重复序列结构（如CCTG、TCTG)，而仅有其中的一种重复序列类型（CCTG）被发现在这种疾病中得到扩展。他们利用人的DNA聚合酶β在体外研究了与这种疾病发生有关的序列相类似的几种重复序列(在重复单位上它们存在一个或几个碱基的差别)，

结果显示，一些细小的变化，如把四碱基重复单位中的T转变成C，可以显著影响新生的重复序列的扩展。因此，作者认为，这种扩展主要是与重复序列的碱基组成有关，而与形成的发夹结构关系不大。由此可见，对于不同的重复序列，其扩展机制也可能不同，或者另有一个共同的机制还没有被揭示出来。

2.2 生物功能与进化

从重复序列在基因的外显子区、内含子以及基因间隔区存在的差异可以看出，串联重复序列在整个基因组中并非随机存在的。位于蛋白质编码区的重复序列，如果发生扩展或减缩，将会因为发生移框突变或因延长的有害 mRNA 的生成，而使基因丧失功能或重新获得新的功能。位于基因非翻译区（untranslated regions，UTRs）5′端重复序列的变异，可以通过影响基因的转录和蛋白质的翻译而起到调节基因表达的作用；位于 UTRs3′端的重复序列的扩展，可以导致转录滑移，并产生延伸的 mRNA，又可以进一步地影响 mRNA 剪切或影响其他的生物功能。同样，位于内含子处的重复序列的扩展或缩减，也可以影响基因转录、mRNA 的剪切或 mRNA 向胞质的输出。位于 UTR 区或者内含子的三碱基重复，还能诱导异染色质样的基因沉默。而最终所有这些影响可以导致生物表型的改变（Li Y et al，2004）。由于位于基因内的比基因间隔区的重复序列往往更具有“生物功能”作用，因而也承受了更多的选择压力，或许是生物快速适应外界环境变化的分子基础。有研究表明，一些生物有机体能快速地改变表型特征来适应各种逆境的环境，就是与基因中的重复序列在 DNA、RNA 或者蛋白质合成过程中产生的滑动错配而诱导产生有关（Rocha et al，2002）。

重复序列的进化主要受以下几个方面因素的影响：重复序列自身的碱基组成情况；重复单位的拷贝数，即重复序列长度；在基因组中的位置；不同物种中重复序列的进化方式和进化速度的差异。

重复序列自身的碱基组成可能决定了重复序列形成过程中的构型，因而必将对其自身的进化产生影响。对小鼠、人类、果蝇和酵母的研究表明，两碱基重复序列类型的突变率最高，其次是三碱基，再次是四碱基（Kruglyak et al，2000）。对于不同重复拷贝类别的两碱基而言，其突变率也是不同的。利用黑腹果蝇（*Drosophila melanogaster*）第二号染色体上 42 个微卫星位点，共包括 3 种重复序列类型（AC、AG、AT），对其 6 个不同群体的突变特征的分析表明，AC 的突变率最高，而 AT 最低（Bachtrog et al，2000）。酵母基因组中含 A+T 的三碱基重复的滑动率较高（Kruglyak et al，2000）。在不同的物种中，各种类型重复序列的滑动突变率可能各不相同（Harr & Schlötterer，2000；Harr et al，2002）。值得注意的是，几乎在所有的生物基因组中，GC 的数目普遍偏少。Schorderet & Gartler（1992）研究了 6 种脊椎动物基因组后，对此作出的解释是：由于基因组 DNA 中的 CpG 的甲基化，使之成为一个突变的热点，因为甲基化的胞苷酸 C 很容易经过脱氨基作用转变成胸腺嘧啶 T。Stallings（1992）的研究结果表明，不管 CpG 两碱基是否受到抑制，物种中的 GC 重复都是偏少的，并进一步认为其可能的问题在于 GC 重复所形成的 DNA 结构上。

重复序列的突变率与其长度成正相关。但重复序列的增长并不是无限的，它取决于滑移突变和序列长度之间的一种平衡。当重复单位的重复拷贝数增加的时候，滑动突变也呈指数式增长。当发生滑动突变的时候，对于短的重复序列而言，倾向于扩展长度；对于长的重复序列，则呈缩减趋势（Lai & Sun，2003）。

重复序列在基因组中的位置不同，其位点多态性组成也不同。这主要是因为承担生物功能的重复序列经受着较大的选择压力，而那些处于“无用”位置的重复序列可以较为“自由”地突变。通过对玉米的一些基因研究表明，67%的启动子、58%的内含子处的重复序列存在多态性，而只有 13%的外显子处的重复序列显示了扩增长度的多态性（Holland et al，2001）。

2.3 进化趋势与 C 值矛盾

Alba 和 Guigo（2004）认为，编码氨基酸的富含 G+C 的重复序列，如 CAG、GAG 等重复可以通过链滑动而得到扩展，这种扩展反过来又可以导致整体基因序列中 G+C 含量的上升；同时，非编码区重复序列中 A 和 T 的比例越高，DNA 双螺旋也越不稳定（Moxon & Rainey，1995）。如果按照这种推论，基因组也将伴随着重复序列的扩增，呈扩大的趋势。这样，越高等的生物，其基因组的含量也应该越大。但事实却存在着一个 C 值矛盾的问题，许多两栖类的基因组比人类的基因组还要大。这也

表明基因组中的重复序列的扩展在高等生物中受到了某种限制。Jiang（1998）以GCG软件和数学模型为工具，对重复序列的宏观组成分析的结果显示，DNA重复序列的存在明显远离平衡态。这种系统地远离平衡态的重复序列的出现，支持重复序列的合成先于基因，基因组起源于重复序列扩增的观点。基因组必须在进化过程中不断形成并维持这种结构，以利基因组发挥正常功能；而不同的重复序列的量都被控制在一定的范围，即为一种耗散结构，是需要消耗能量来维持的，因而有它特定的功能。在进化过程中，基因组为了对抗环境压力需增加新的功能和新的基因。这种增加可以从增加重复序列开始，重复序列可进一步突变或为某些序列的扩增或移动提供重组所需的相同序列；同时，过多的重复序列将导致能量消耗的过重负担以及过多的重组和剪切，即导致基因组的不稳定性。因此，重复序列必须被维持在某一范围内。

通过以上的分析，我们认为C值矛盾看似矛盾，实则根本不矛盾。基因组扩增的一个重要来源是重复序列的扩增，而生物进化的目标并不体现在生物基因组的庞大上，而是体现在基因组的“效率”上。过多的重复序列的扩增，相反的是生物机体内与DNA合成修复等相关酶系统不完善的体现，将导致大量的无用序列的出现，这些序列必将降低基因组发挥功能的效率，是与“进化”本意相违背的。而只有像人类等这种高等生物会把基因组的规模控制在一个较为完善的水平，而过低的基因组水平，如大肠杆菌等原核生物，虽然基因组发挥生物功能的效率最高，但其所能表现的生物功能却很有限。

3 问题与展望

目前对于重复序列在物种基因组中的组成与存在特征等研究，还存在着研究方法和技术上的差别。这些差别给不同生物基因组间的比较研究带来了许多麻烦，同时也产生了许多新的研究课题。

3.1 重复序列分析软件的多样化

目前可用于重复序列分析的软件很多，如Tandem repeat finder、Repeat masker、DNA works等。由于研究者采用的重复序列分析软件和算法各不相同，这给不同生物基因组间的重复序列在组成和存在特征上的比较带来困难，甚至不同软件对同一物种中重复序列分析结果也不尽相同。如Klevytska et al（2001）用Genequest软件程序（Dnastar package，LaserGene Inc. Madison Wis.，分析重复序列的起点长度是8个碱基，可以对1～100 bp以上的重复单位组成的重复序列进行分析）与用SSR Search程序（可以对小至两个碱基长度的重复序列进行分析，但只能对1～10个碱基长度的重复单位组成的重复序列进行分析）对鼠疫耶尔森氏菌（*Y. pestis*）基因组中的重复序列分析结果做了比较，两者的结果明显不同：前者三碱基的数量最多，其次是六碱基和八碱基；而后者单碱基最多，然后依次分别是二碱基、三碱基、四碱基等。SSR Search在*Y. pestis*基因组中，每一万碱基序列长度可以检测到950个SSR，而Genequest只能发现1.86个SSR。分析结果差异源于对重复序列搜索的算法不同。Genequest可以检测不完美的重复序列，同时最小检测8个碱基长度；而SSR Search搜索程序可以搜索许多前者所不能搜索到的重复序列，但却只能搜索1～10个碱基重复单位组成的重复序列（大量短的单碱基），同时也不能检测不完美型的重复序列。

3.2 重复序列统计标准的差异

重复序列统计标准的差异也给不同生物基因组之间的比较带来困难。其差异主要有3个方面：重复序列的长度定义、重复序列完美性（perfect）的程度、DNA双链中重复序列的互补情况。

3.2.1 重复序列的长度定义　Klevytska et al（2001）把8个及其以上的碱基长度作为统计标准，即：对于单碱基重复，要连续8个；两碱基的重复拷贝数为4个或4个以上；三碱基的重复拷贝数为3个或3个以上，依次类推。Rocha et al（2002）对此的定义则是：≥5的单核苷酸序列，≥6的两核苷酸序列，≥6的三核苷酸序列，≥8的四核苷酸序列，即把最小的重复长度定义为5个碱基。此外，还有把12个碱基的长度（Borstnik & Pumpernik，2002），甚至20个碱基的长度（Katti et al，2001）作为统计的起点。同时，也有以最低重复拷贝数作为统计的标准，如Ross et al（2003）把最低拷贝数定为5，即5个碱基长度的单碱基重复序列类型，10个碱基长度的两碱基重复序列类型，依次类推。这些统计标准的差异对不同文献中所报道的不同物种基因组间重复序列的比较，同时也对同一物种中的分析结果产生歧异。

我们认为，由于重复序列在基因组中存在的随机性和必然性，对重复序列统计标准的界定，既要

考虑重复序列的长度，也要考虑重复序列的拷贝数。就随机性而言，在DNA序列4种碱基中，随意两个相同碱基并列在一起的几率远大于3个相同碱基，这种差异只是反映了碱基随机排列出现重复序列的概率；而就必然性而言，重复序列的发生是基因组长期进化的产物，在基因组中是有一定生物学功能的，因此，不能把重复序列的长度定义得太小。如对单碱基的定义，如果只以2或3个碱基长度作为统计起点，这种类型无疑是在所有基因组中最丰富的重复序列类型，掩盖了重复序列发生的“必然性”的本质。当然，长度越长，也就越“严格”，但会漏掉大量的有用信息。因此，我们认为把重复序列最低长度定义在10～20个碱基，同时对于七碱基及其以上重复序列类型的拷贝数定义在2个拷贝较为合适。因而具体的统计标准是：14或14个拷贝以上的单碱基重复序列，7或7个拷贝以上的两碱基重复序列；5或5个拷贝以上的三碱基重复序列；4或4个拷贝以上的四碱基重复序列，3或3个拷贝以上的五六碱基重复序列（Gao et al, 2004)，2或2个拷贝以上的七碱基及七碱基以上的重复序列类型（Gao & Kong，2005)。

3.2.2 重复序列完美性的程度 Weber（1990）针对人类（CA)n中CA重复序列排列方式的不同，提出两碱基重复序列类型的微卫星排列方式可以划分为3种：完美型（perfect)、不完美型（imperfect）和复合型（compound)。完美型是指核心序列以不间断的重复方式首尾相连而成；不完美型是指2个或2个以上的同种重复序列被3或3个以下的非重复碱基所间隔；复合型是指一种重复序列和其他种重复序列由3个碱基以下的非重复序列间隔（包括直接相连接）所组成的重复序列类型。有些以完美型作为统计标准（Tóth et al，2000；Temnykh et al，2001)；另一些则以不完美型作为统计标准，如Katti et al（2001）允许每10个碱基可以有1个碱基“错配”，而Rose et al（2003）采用的统计标准则更为宽松。由于各自采用的标准不同，会对统计结果产生很大的影响。

3.2.3 DNA双链中重复序列的互补情况 Gur-Arie et al（2000）统计两碱基重复时，认为有6种类型，即AC/CA、AG/GA、AT/TA、CG/GC、CT/TC、GT/TG。而事实上，AC/CA和GT/TG、AG/GA和CT/TC只是反应了DNA互补链的不同，本质上属于同一种类型。在考虑互补链的同时，还应该考虑计数重复拷贝数起始顺序的差异。如ATATATATATATAT，可以看作是AT重复单位重复7次所组成的重复序列，也可以看作TA重复单位重复6.5次所组成的重复序列。因此，应该把AT和TA重复看作同一种类型，虽然这会造成重复拷贝数上0.5个拷贝数的差异，但这对各重复序列类型中的重复拷贝数存在特征影响很小。

3.3 展望

随着DNA测序技术的改进，近几年来，DNA测序进程大大加快，每天都有大量的核酸序列提交到GeneBank等数据库上，这有利于系统分析各物种基因组中重复序列的组成和存在。今后，串联重复序列研究的热点将集中在以下几个方面：

3.3.1 串联重复序列自身的进化关系 目前的大量研究均限于微卫星等（Katti et al，2001；Tóth et al，2000)，而对于六核苷酸重复单位以上的重复序列在基因组中的组成和存在研究较少，且仅限于一些原核生物（Klevytska et al，2001)；很少涉及真核生物，尤其是重复单位长度在20个碱基以上的重复序列几乎处于空白。其中的原因在于：对更长重复单位组成的重复序列的研究，会使其统计工作量成倍地增加；Guo（2004）认为：更长的微卫星（在5个碱基重复单位长度以上）在表现规律上没有更多的变化，增加微卫星重复单位的长度只是增加计算量，不会得到更多的结果。但对中国明对虾基因组的研究结果表明，微卫星和小卫星重复序列在发生上可能存在一定联系，即一部分小卫星可能是在微卫星重复序列的基础上进化而来的（Gao & Kong，2005)。因此，随着微卫星研究基础资料的积累以及对于更长重复单位组成的重复序列的调查分析，这个问题有望在不久的将来得到明确的答案。扩大对于小卫星重复序列在基因组中，尤其是在人类、果蝇等基因组序列已经测序完毕的生物中的调查分析是非常有必要的，这也是将来重复序列研究分析的一个重要方向。

3.3.2 串联重复序列在全基因组水平上的进化地位 重复序列一度曾被认为是进化中的垃圾序列，曾被称为进化的痕迹。而现在普遍的看法是，许多重复序列在基因组中具有显著的生物功能。人类基因组中的微卫星比例为3%（International Human Genome Sequencing Consortium，2001)，按蚊基因组中微卫星约占整个基因组序列的2.14%（Yu et al，2005)，而河豚鱼（*Takifugu rubripes*）中的约为

0.96%（Takagi et al，2003）。相反，在我们研究的中国明对虾基因组中，微卫星的比例则高达9.78%，小卫星约占3.42%，串联重复序列总体上占整个基因组的13.2%（Kong & Gao，2005）。同时通过对果蝇属的5个物种（*D. arizonae*、*D. mojawensis*、*D. pachea*、*D. neatestacea* 和 *D. recens*）基因组中的微卫星的组成与存在等特征的研究表明，微卫星在这5个种间的存在与组成方面的差异较少（Ross et al，2003）。由此可见，不同生物中串联重复序列在基因组中的组成比例是不同的。而亲缘关系相近的物种，串联重复序列在基因组中的组成和存在又具有一定的相似性，这显示了其与生物基因组进化上的密切关系。如果基于系统进化树中各种生物的进化关系研究各种串联重复序列在物种中的组成与存在特征，将会有许多新的发现。这也就是新的课题：基于串联重复序列进化关系的比较基因组学研究。

3.3.3 串联重复序列在基因组中的功能 越来越多的研究表明，许多串联重复序列在基因组中具有重要的功能，目前认为最起码具有3个作用：一是组成开放阅读框的一部分；二是参与基因组的调节活动；三是组成染色体的脆性位点（Vergnaud & Denoeud，2000）。前面已经提到许多三碱基重复序列直接与许多遗传疾病的发生密切相关（Margolis & Ross，2001），其主要原因就是这些重复序列与其所在染色体的脆性位点有关。有趣的是，在性别决定的研究中，还发现了GATA重复序列参与了性别调控。如，这类重复序列在蛇的性染色体的进化和分化中起着重要作用；在真核类生物中，还普遍与性别决定染色体密切有关（Subramanian et al，2003）。还有些类似小卫星重复序列的DNA序列，甚至是直接用来编码蛋白质的（Marinangeli et al，2004）。可以说，对串联重复序列生物功能的研究还处于起步阶段，而这项研究与生物基因组的进化研究是密切相关的，因而也必将成为一个新的研究热点。

3.3.4 重复序列数据库的建立和完善 串联重复序列在基因组中的组成与存在等方面的研究是生物信息学在基因组研究中的重要研究领域。目前利用生物信息学手段对重复序列的研究在国外蓬勃开展，而国内则处于刚刚起步的阶段。在这方面最显著的成就是建立了部分生物基因组串联重复序列数据库，如 Vergnaud & Denoeud（2000）已经建立了包括人类、秀丽隐杆线虫、拟南芥和一些原核生物的串联重复序列搜索数据库（http：//minisatellites.u-psud.fr）。而 Katti et al（2001）也建立了人的21和22号染色体、大果蝇（*D. melanogaster*）、秀丽隐杆线虫、拟南芥和酵母中的微卫星重复序列数据库，并对各个重复序列在染色体中的位置作了详尽的注释(http://www.ncl-india.org/ssr)。同时更多的重复序列信息数据也可以在网上得到(ftp://ftp.technion.ac.il/pub/supported/biotech/ssr.exe, Klevytska et al，2001；http://genetics.elte.hu/ssr; Tóth et al，2000)。随着我国一些特有的珍惜生物(如大熊猫、大鲵）和一些特有的水产养殖生物(如中国明对虾等）基因组的研究深入，相应的串联重复序列数据库将会得以建立。

3.3.5 串联重复序列的应用研究 目前微卫星遗传标记作为一类稳定而常规化的遗传分析技术，已经在包括人类的亲子鉴定、各种动植物的家系分析、遗传多样性、遗传连锁作图、疾病的遗传连锁分析等方面得到了应用。另一种基于微卫星重复序列的遗传标记是区间简单重复序列（inter simple sequence repeat，ISSR）技术，也是一种常用的遗传标记。这两种遗传标记都是基于PCR技术的。另外基于Southern杂交的小卫星DNA指纹图谱技术，自Jeffreys（1985）创立以来，也得到了广泛地应用。如，利用短片段寡核苷酸探针（微卫星重复序列）的指纹图谱分析也同样可以得到较好的遗传指纹图谱（Liu et al，2000）。随着这些遗传标记的开发和研究，必将促进育种学中的数量性状基因定位和疾病筛查等与人类自身利益密切相关的学科发展。

总之，随着生物信息学和基因组学等的研究，我们对串联重复序列在基因组中的生物功能作用的认识正在逐渐加深；通过这些重复序列在不同物种间和物种内的比较分析，将有助于了解基因组的起源和进化，也会更好地发挥这些重复序列在基因表达调节、群体遗传多样性分析，以及开发分子标记等方面的作用，并得到更为广泛地应用。

参考文献：

Alba MM, Guigo R. 2004. Comparative analysis of amino acid repeats in rodents and humans [J]. *Genome Res*, **14**: 549 - 554.

Bachtrog D, Agis M, Imhof M, Schlötterer C. 2000. Microsatellite variability differs between dinucleotide repeat motifs: Evidence from *Drosophila melanogaster* [J]. *Mol Biol Evol*, **17**: 1277 - 1285.

Bachtrog D, Weiss S, Zangerl B, Schlötterer C. 1999. Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome [J]. *Mol Biol Evol*, **16**: 602 - 610.

Beckman JS, Weber JL. 1992. Survey of human and rat microsatellites [J]. *Genomics*, **12** (4): 627 - 631.

Belkum AV, Scherer S, Alphen LV, Henri V. 1998. Short-sequence DNA repeats in prokaryotic genomes [J]. *Microbiol Mol Biol Rev*, **62** (2): 275 - 293.

Borstnik B, Pumpernik D. 2002. Tandem repeats in protein coding regions of primate genes [J]. *Genome Res*, **12** (6): 909 - 915.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. 1998. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat [J]. *Am J Hum Genet*, **62**: 1408 - 1415.

Cox R, Mirkin SM. 1997. Characteristic enrichment of DNA repeats in different genomes [J]. *Proc Natl Acad Sci USA*, **94**: 5237 - 5242.

Edwards YJ, Elgar G, Clark MS, Bishop MJ. 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: Perspectives in functional and comparative genomic analyses [J]. *J Mol Biol*, **278**: 843 - 854.

Gao H, Kong J. 2005. An analysis of minisatellite repeat sequences in Chinese shrimp (*Fenneropenaeus chinensis*) genome [J]. *Acta Zool Sin*, **51** (1): 101 - 107. [高 焕, 孔 杰. 2005. 中国明对虾基因组小卫星重复序列分析. 动物学报, **51** (1): 101 - 107.]

Gao H, Liu P, Meng XH, Wang WJ, Kong J. 2004. Analysis of microsatellite sequences in Chinese shrimp (*Fenneropenaeus chinensis*) genome [J]. *Oceanol Limnolo Sin*, **35** (5): 424 - 431. [高 焕, 刘 萍, 孟宪红, 王伟继, 孔 杰. 2004. 中国对虾 (*Fenneropenaeus chinensis*) 基因组微卫星特征分析. 海洋与湖沼, **35** (5): 424 - 431.]

Guo WJ. 2004. Primary Research on the Microsatellite Distribution and Function in Genomics and the Relevant Computational Methodology [D]. Ph. D. thesis, Sichuan Agriculture University. [郭文久. 2004. 微卫星在基因组上的分布与功能及其计算方法初步研究. 四川农业大学博士学位论文.]

Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y. 2000. Simple sequence repeats in *Escherichia coli*: Abundance, distribution, composition, and polymorphism [J]. *Genome Res*, **10** (1): 62 - 71.

Hadonou AM, Sargent DJ, Wilson F, James CM, Simpson DW. 2004. Development of microsatellite markers in *Fragaria*, their use in genetic diversity analysis, and their potential for genetic linkage mappi [J]. *Genome*, **47** (3): 429 - 438.

Harr B, Schlötterer C. 2000. Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation [J]. *Genetics*, **155**: 1213 - 1220.

Harr B, Todorova J, Schlötterer C. 2002. Mismatch repair driven mutational bias in *D. melanogaster* [J]. *Mol Cell*, **10**: 199 - 205.

He P. 1998. Abundance, polymorphism and applications of microsatellite in eukaryote [J]. *Hereditas* (*Beijing*), **20** (4): 42 - 47. [何 平. 1998. 真核生物中的微卫星及其应用. 遗传, **20** (4): 42 - 47.]

Heidenfelder BL, Makhov AM, Topal MD. 2003. Hairpin formation in Friedreich's ataxia triplet repeat expansion [J]. *J Biol Chem*, **278** (4): 2425 - 2431.

Heidenfelder BL, Topal MD. 2003. Effects of sequence on repeat expansion during DNA replication [J]. *Nucleic Acids Res*, **31** (24): 7159 - 7164.

Holland JB, Helland SJ, Sharopova N, Rhyne DC. 2001. Polymorphism of PCR-based markers targeting exons, introns, promoter regions, and SSRs in maize and introns and repeat sequences in oat [J]. *Genome*, **44** (6): 1065 - 1076.

Ingavale SS, Kaur R, Aggarwal P, Bachhawat AK. 1998. A minisatellite sequence within the propeptide region of the vacuolar carboxypeptidase Y gene of *Schizosaccharomyces pombe* [J]. *J Bacteriol*, **180** (14): 3727 - 3729.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome [J]. *Nature*, **409**: 860 - 921.

Jauert PA, Edmiston SN, Conway K, Kirkpatrick DT. 2002. RAD1 controls the meiotic expansion of the human *HRAS*1 minisatellite in *Saccharomyces cerevisiae* [J]. *Mol Cell Biol*, **22** (3): 953 - 964.

Jeffreys AJ, Wilson V, Thein SL. 1985. Hypervariable' minisatellite' regions in human DNA [J]. *Nature*, **314** (6006): 67 - 73.

Jiang H. 1998. The distribution trends in simple repetitive stretches of DNA [J]. *Chinese J Biochem Mol Biol*, **14** (1): 65 - 70. [江 洪. 1998. DNA重复序列的宏观分布趋势. 中国生物化学与分子生物学报, **14** (1): 65 - 70.]

Katti MV, Ranjekar PK, Gupta VS. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences [J]. *Mol Biol Evol*, **18**: 1161 - 1167.

Klevytska AM, Price LB, Schupp JM, Worsham PL, Wong J, Keim P. 2001. Identification and characterization of variable-number tandem repeats in the *Yersinia pestis* genome [J]. *J Clin Microbiol*, **39** (9): 3179 - 3185.

Kong J, Gao H. 2005. Analysis of tandem repeats in the genome of Chinese shrimp *Fenneropenaeus chinensis* [J]. *Chinese Science Bulletin*, **50** (14): 1462 - 1469.

Kruglyak S, Durrett R, Schug MD, Aquadro CF. 2000. Distribution and abundance of microsatellites in the yeast genome can be explained by a balance between slippage events and point mutations [J]. *Mol Biol Evol*, **17**: 1210 - 1219.

Lai Y, Sun F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units [J]. *Mol Biol Evol*, **20** (12): 2123 - 213.

Levinson G, Gutman GA. 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution [J]. *Mol Biol Evol*, **4**: 203 - 221.

Li B, Xia Q, Lu C, Zhou Z, Xiang Z. 2004. Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes [J]. *Geno Prot Bioinfo*, **2** (1): 24 - 31.

Li Y, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: Structure, function, and evolution [J]. *Mol Biol Evol*, **21** (6): 991 - 1007.

Liu J, Liu GS, Qi DM, Li FF. 2000. Construction of genetic fingerprints of *Aneurolepidium chinensis* using microsatellite sequences [J]. *Acta Bot Sin*, **42** (9): 985 - 987. [刘 杰, 刘公社, 齐冬梅, 李芳芳. 2000. 用微卫星序列构建羊草遗传指纹图谱. 植物学报, **42** (9): 985 - 987.]

Luo WY, Hu J, Li XF. 2003. The evolution and application of microsatellites [J]. *Hereditas*, **25** (5): 615 - 619. [罗文永, 胡骏, 李晓方. 2003. 微卫星序列及其应用. 遗传, **25** (5): 615 - 619.]

Margolis RL, Ross CA. 2001. Expansion explosion: New clues to the pathogenesis of repeat expansion neurodegenerative diseases [J]. *Trends Mol Med*, **7**: 479 - 482.

Marinangeli P, Angelozzi D, Ciani M, Clementi F, Mannazzu I. 2004. Minisatellites in *Saccharomyces cerevisiae* genes encoding cell wall proteins: A new way towards wine strain characterisation [J]. *FEMS Yeast Res*, **4** (4-5): 427-435.

Moxon ER, Rainey PB. 1995. Pathogenic bacteria: The wisdom of their genes [A]. In: Van der Zeijst BAM, Van Alphen L, Hoekstra WPM, van Embden JDA . Ecology of Pathogenic Bacteria. Second Series. No. 96 [M]. Amsterdam: Royal Dutch Academy of Sciences, 255-268.

Rocha EP, Matic I, Taddei F. 2002. Over-representation of repeats in stress response genes: A strategy to increase versatity under stressful conditions [J]. *Nucleic Acids Res*, **30** (9): 1886-1894.

Romero C, Pedryc A, Munoz V, Llacer G, Badenes ML. 2003. Genetic diversity of different apricot geographical groups determined by SSR markers [J]. *Genome*, **46** (2): 244-252.

Ross CL, Dyer KA, Erez T, Miller SJ, Jaenike J, Markow TA. 2003. Rapid divergence of microsatellite abundance among species of *Drosophila* [J]. *Mol Biol Evol*, **20** (7): 1143-1157.

Saha A, Bamezai R. 2000. Detection of genetic variation in Indian population groups using a novel minisatellite probe and finding relationships through tree construction [J]. *J Hum Genet*, **45** (4): 207-211.

Sakurai K, Horiuchi Y, Ikeda H, Ikezaki K, Yoshimoto T, Fukui M, Arinami T. 2004. A novel susceptibility locus for moyamoya disease on chromosome 8q23 [J]. *J Hum Genet*, **49** (5): 278-281.

Schorderet DF, Gartler SM. 1992. Analysis of CpG suppression in methylated and nonmethylated species [J]. *Proc Natl Acad Sci USA*, **89**: 957-961.

Selvamani MJ, Degnan SM, Degnan BM. 2001. Microsatellite genotyping of individual abalone larvae: Parentage assignment in aquaculture [J]. *Mar Biotechnol* (NY), **3** (5): 478-485.

Stallings RL. 1992. CpG suppression in vertebrate genomes does not account for the rarity of (CpG)n microsatellite repeats [J]. *Genomics*, **13**: 890-891.

Staten R, Schully SD, Noor MA. 2004. A microsatellite linkage map of *Drosophila mojavensis* [J]. *BMC Genet*, **5** (1): 12.

Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of Bkm sequences (GATA repeats): Predominant association with sex chromosomes and potential role in higher order chromatin organization and function [J]. *Bioinformatics*, **19** (6): 681-685.

Takagi M, Sato J, Monbayashi C, Aoki K, Tsuji T, Hatanaka H, Takahashi H, Harumi S. 2003. Evaluation of microsatellites identified in the tiger puffer *Takifugu rubripes* DNA database [J]. *Fisheries Sci*, **69**: 1085-1095.

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential [J]. *Genome Res*, **11** (8): 1441-1452.

Thibodeau SN, Bren G, Schaid D. 1993. Microsatellite instability in cancer of the proximal colon [J]. *Science*, **260** (5109): 816-819.

Tóth G, Gáspári Z, Jurka J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis [J]. *Genome Res*, **10** (7): 967-981.

Vazquez F, Perez T, Albornoz J, Domingue Z. 2003. Estimation of the mutation rates in *Drosophila melanogaster* [J]. *Genet Res*, **76**: 323-326.

Vergnaud G, Denoeud F. 2000. Minisatellites: Mutability and genome architecture [J]. *Genome Res*, **10** (7): 899-907.

Weber JL. 1990. Informativeness of human (dC-dA)n· (dG-dT)n polymorphisms [J]. *Genomics*, **7** (4): 524-530.

Wickstead B, Ersfeld K, Gull K. 2003. Repetitive elements in genomes of parasitic protozoa [J]. *Microbiol Mol Biol Rev*, **67** (3): 360-375.

Xu X , Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length [J]. *Nat Genet*, **24** (4): 396-399.

Yang J, Wang J, Chen L, Yu J, Dong J, Yao ZJ, Shen Y, Jin Q, Chen R. 2003. Identification and characterization of simple sequence repeats in the genomes of *Shigella* species [J]. *Gene*, **322**: 85-92.

Yu QY, Li B, Li GR, Fang SM, Yan H, Tong XL, Qian JF, Xia QY, Lu C. 2005. Abundance and distribution of microsatellites in the entire mosquito genome [J]. *Prog Biochem Biophys*, **32** (5): 435-441. [余泉友, 李 斌, 李关荣, 房守敏, 颜 虹, 童晓玲, 钱吉凤, 夏庆友, 鲁 成. 2005. 蚊子全基因组中微卫星的丰度及其分布. 生物化学与生物物理进展, **32** (5): 435-441.]